NIFA Enterprise Search 1.0 User Guide

Last updated January 6, 2014

Prepared for

The National Institute of Food and Agriculture

NIFA

By

Barquin International

1707 L St, N.W., Suite 570
Washington, D.C 20006
202-296-7147

Table of Contents

1 Background

This document serves as a brief introduction to the functionality available in Phase 1 (Release 1.0) of the NIFA Enterprise Search module that replaces the existing Google Search Appliance (GSA) for both the REEIS public website (http://reeis.usda.gov) and the Leadership Management Dashboard (LMD).

The primary major goal for Phase 1 was to leverage the existing data extraction, transformation, and loading (ETL) processes used by the GSA, and to provide added features not available using the GSA, such as saving and sharing search definitions and snapshots of search results.

In subsequent releases we will add additional data elements and additional search domains with guidance from NIFA management and stakeholders.

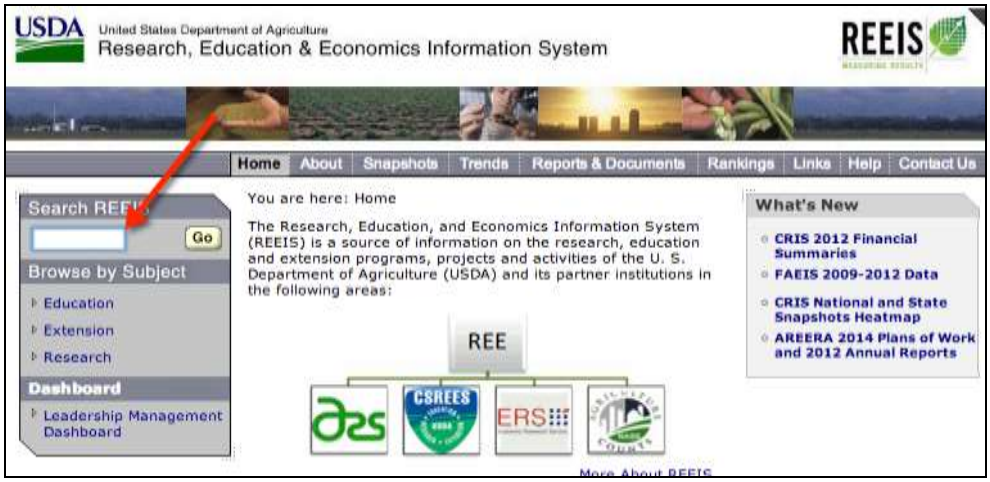
2 Brief Overview

2.1 How to Access the NIFA Enterprise Search

After logging into LMD, select the "Search" menu item:

This is the view of the Search menu from the LMD 4.5 version of the menus:

From the REEIS public website use the "Search REEIS" box as before:



2.1.1 Search Domains

The following four search domains are available in Phase 1. (Additional domains are planned for subsequent releases.)

1. Project Components (Faceted)
    - Data Source: CRIS (now REEport) and C-REEMS, updated nightly.
    - Formerly the CRIS Faceted Search (searches each text field individually like the Objectives for example)
2. Project Pages
    - Data Source: CRIS (now REEport), updated nightly.
    - Formerly CRIS Projects Search (searches entire project instead of each field separately)
3. Annual Reports on Planned Programs
    - Data Source: POW (Plan of Work System), updated nightly, but major changes are seasonal.
    - Formerly POW Annual Reports on Planned Programs
4. REEIS Website
    - Data Source: http://reeis.usda.gov, updated nightly, but changes are infrequent.

Note that although casual users can search by entering terms and phrases easily using the Boolean "AND" and "OR" operators, the full capabilities of the Lucene query syntax is supported in each of the search domains. The next release of NIFA Enterprise Search will provide a "wizard" for creating and using more advanced queries.

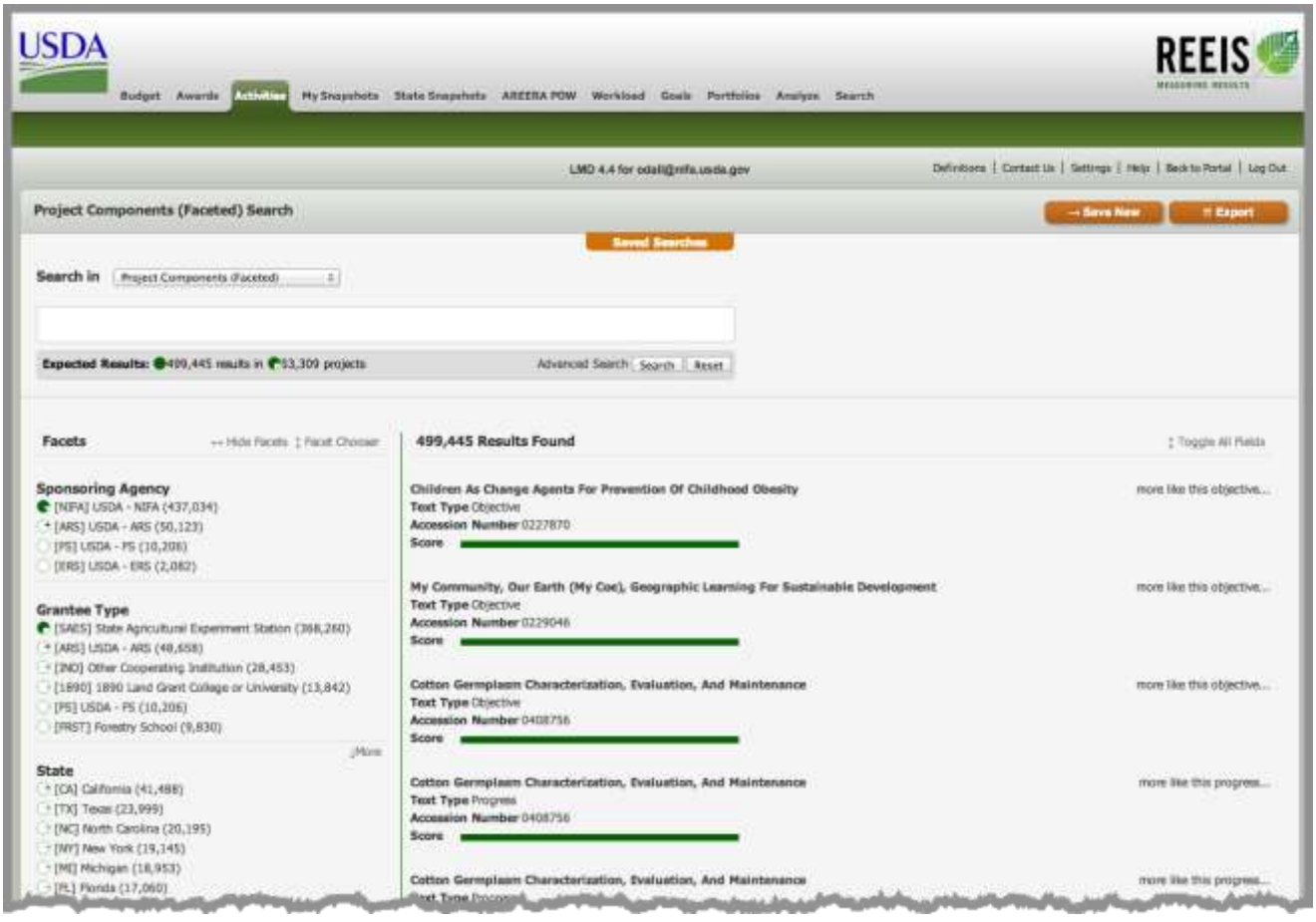
OIT will soon post introductory screencasts that will be available from the "help" link:

2.1.2 Project Components

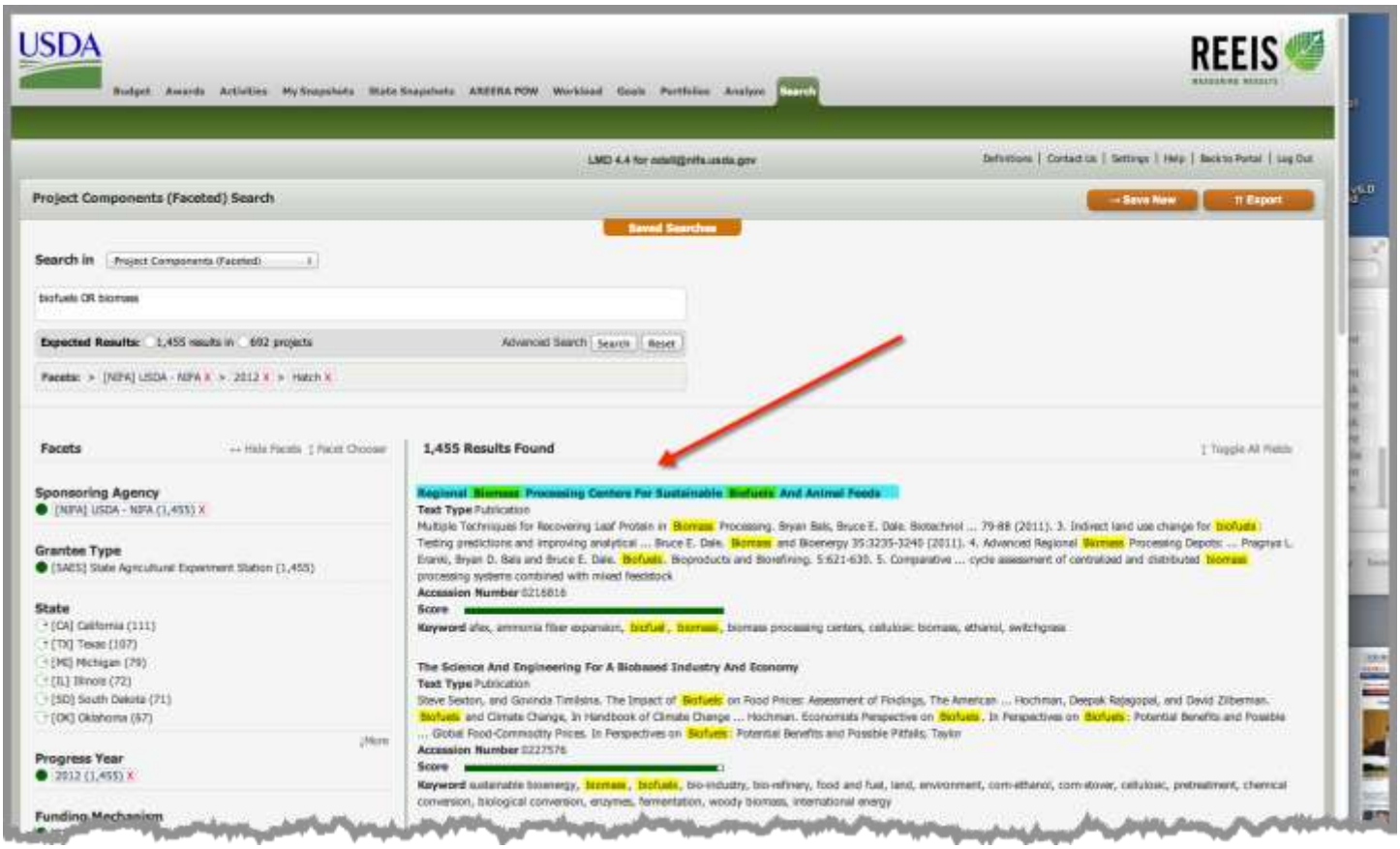
This was formerly called "CRIS Faceted Search"

The Project Components (Faceted) search is the most detailed and powerful of the four initial search domains. It includes a large number of CRIS (now REEport) project fields integrated with NIFA financial data from C-REEMS (FDC, Grant Year, Award Amount Ranges). Project Components search provides the integration of a powerful text search engine (Lucene) with the addition of filtering using pre-defined categories or "Facets" that provide a deterministic drill-through "bread crumb" trail that includes a hit count for each step through the combination of facet choices. Facets available in Phase 1 include:

1. Sponsoring Agency
2. Grantee Type
3. State
4. Grant Award Year
5. Progress Year
6. Funding Mechanism
7. Financial Data Code (FDC)
8. Keyword
9. Knowledge Area
10. Subject of Investigation
11. Field of Science
12. Text Field Type
13. Award Amount Range



Each "Project Component" (Objectives, Progress, Impact, Publications) can be investigated and exported independently from the related project. Thus, one can focus a search, for example, on the Impact statements reported for the Progress Year of 2012 for Hatch projects that have the terms "biofuel" or "biomass" somewhere in one of the project fields:



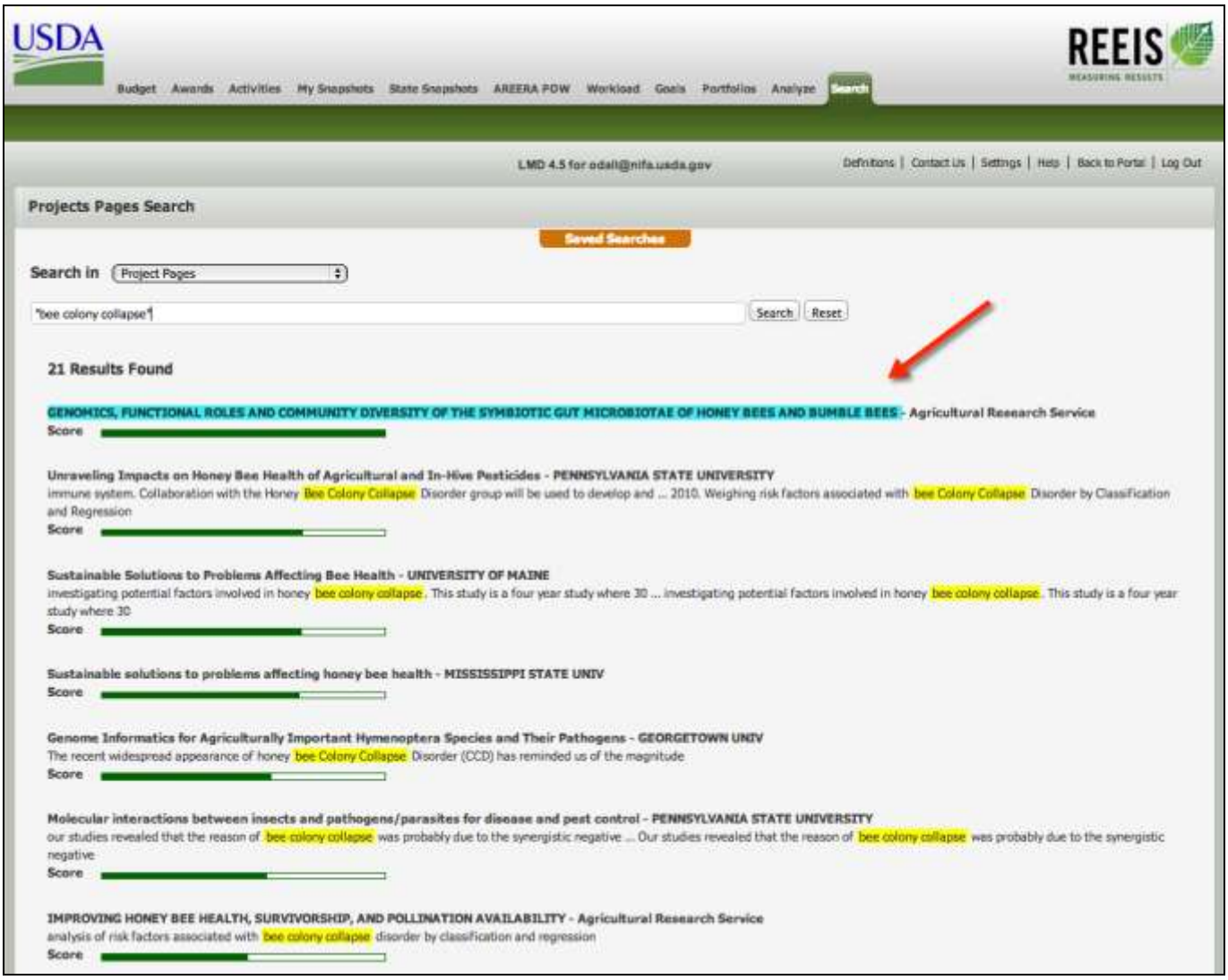
Currently Sponsoring Agencies included in this domain are NIFA, ARS, ERS, and FS (Forest Service). (In the next release it has been recommended to add State Partner sponsored projects as well.)

Clicking on a hyperlinked title (see above screen shot) in the search results will launch a view of the entire project detail page.

2.1.3 Project Pages

This was formerly called "CRIS Projects"

This search domain is built by indexing the each entire project detail pages (HTML) for all available CRIS (now REEport) projects. The search interface appears as follows:



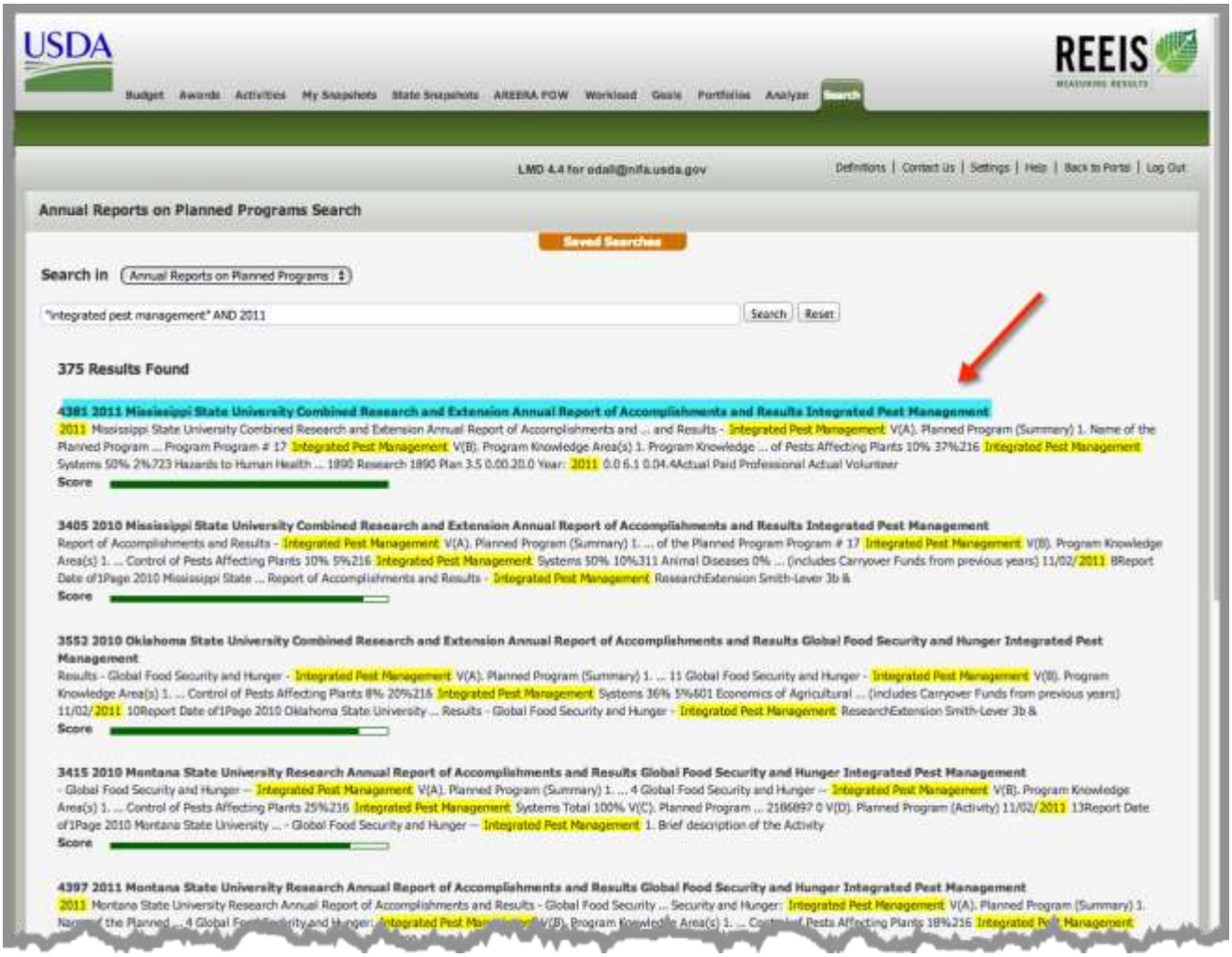
In contrast to the Project Faceted Search, which indexes project text components separately, each project summary is indexed as one document, without facets.

Clicking on a hyperlinked title (see above screen shot) in the search results will launch a view of the entire project detail page.

2.1.4 Annual Reports on Planned Programs

This was formerly called "POW Annual Reports on Planned Programs"

The search interface appears as follows:

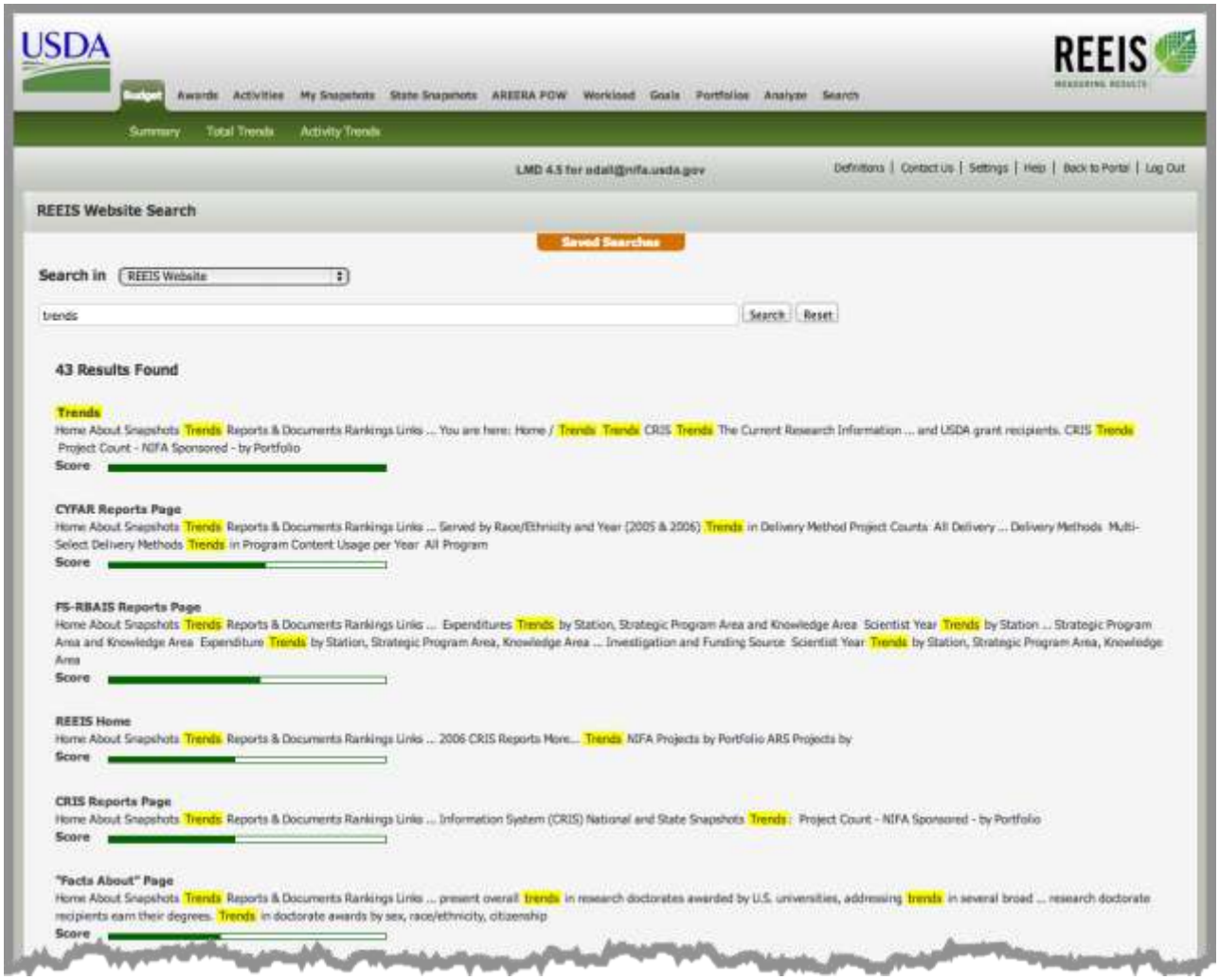


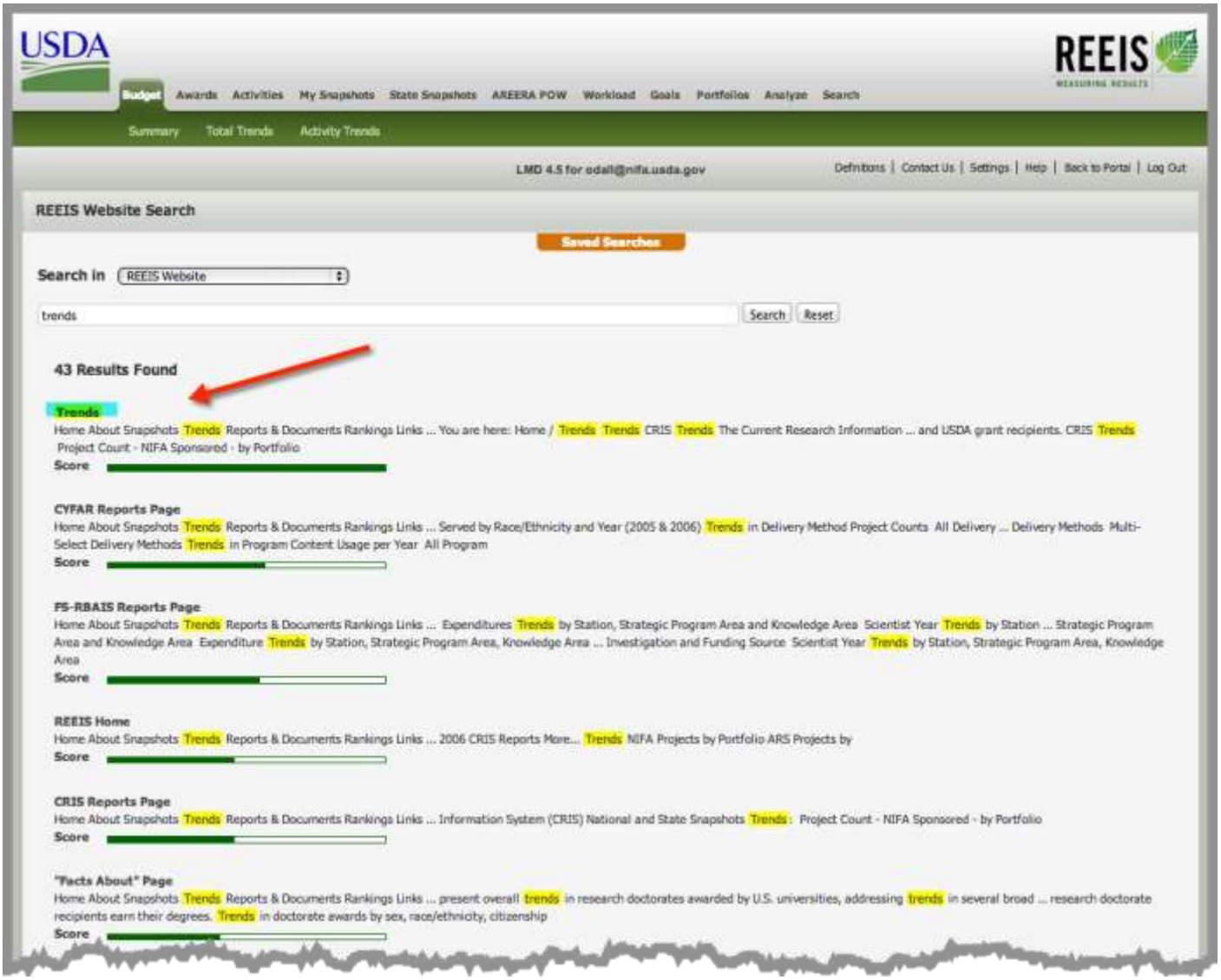
Clicking on a hyperlinked title (see above screen shot) in the search results will launch a view of the entire Annual Report of Accomplishments for the Planned Program.

2.1.5 REEIS Website

This was formerly called "REEIS Portal"

The search interface appears as follows:





Clicking on a hyperlinked title (see above screen shot) in the search results will launch a view of the entire REEIS Website page found in the search results.

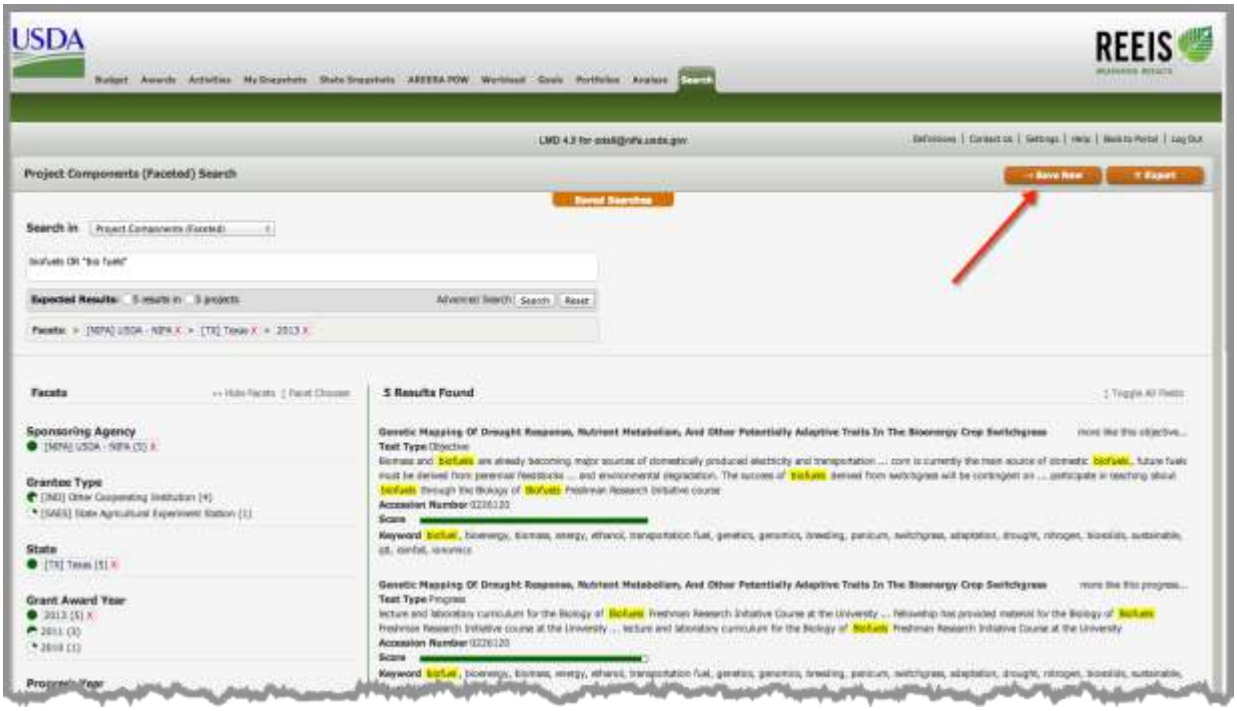
2.2 Saving and Exporting Search Results

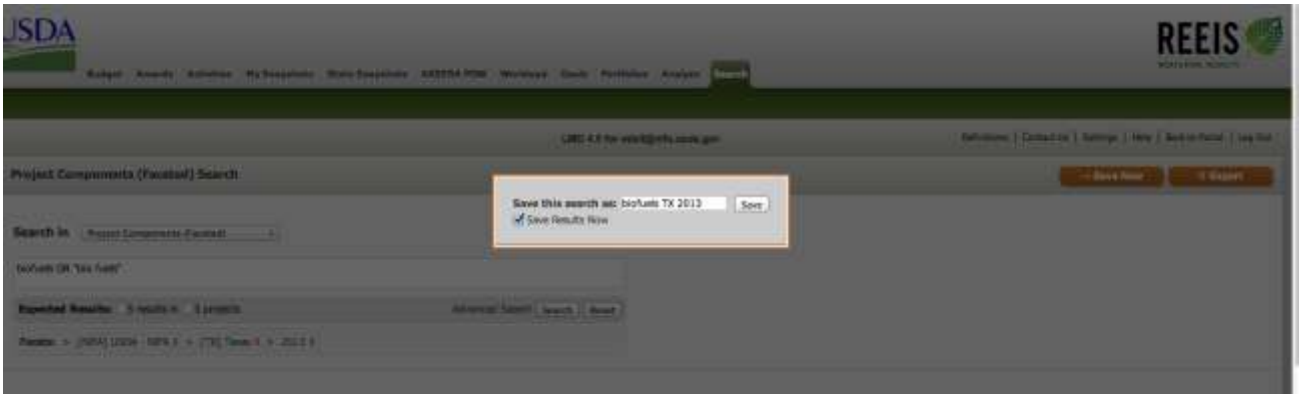
2.2.1 Definitions

Saved Search. A "Saved Search" is a saved search definition with a user-specified name (title), which includes the text specified as well as the facets selected during the search. However, every time the search is loaded, fresh results are returned according to the latest data/index.

Saved Result. A "Saved Result" is a snapshot in time of the data that are in the data/index at the time of query execution. Each "Saved Search" can have many related snapshots.

2.2.2 Saving a Search Definition



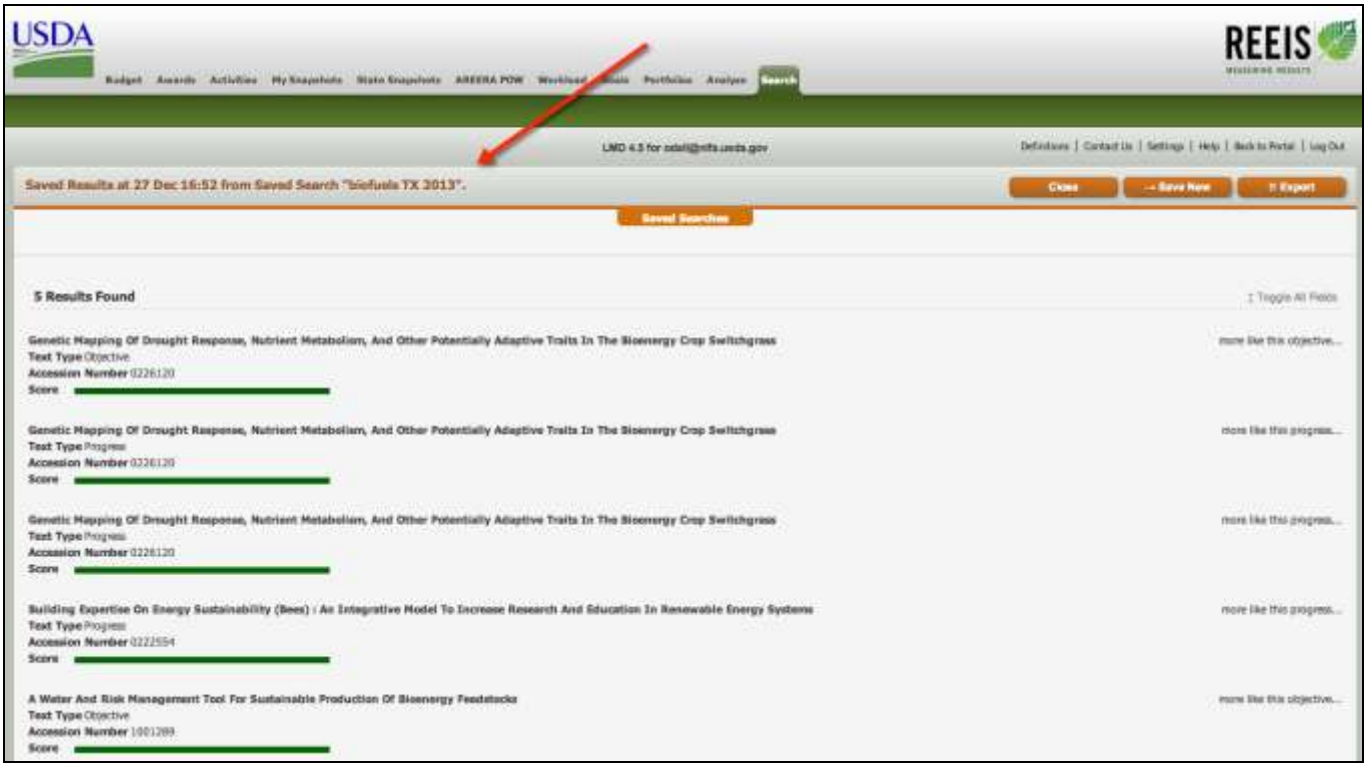


Note that also selecting the "Save Results Now" option will create a snapshot of data that exists at that time using that saved search definition.

Clicking on the "Saved Searches" button will display the current list of saved searches in the LMD repository:

Clicking on "Results Panel" related to a Saved Search will bring up the summary of Saved Results for this query (if it exists) because we chose that option while saving the search definition. Note that only the first item in the Saved Searches List has the options of "Save Results" and "Delete Search" as it is the only saved search owned by me (Owen Dall). You can only delete and modify searches that you have created. Clicking on the "Results Panel" will display the following.

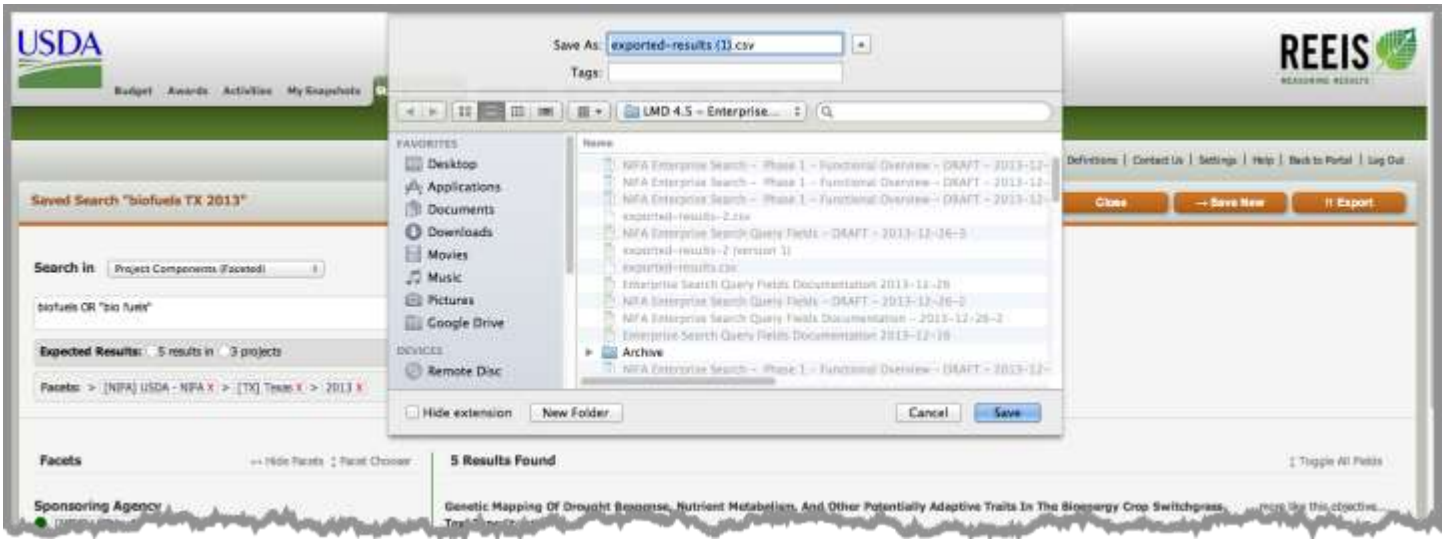
Clicking on "Show Results" will display the following:


USDA
REEIS
Budget
Awards
Activities
My Snapshots
State Snapshots
AREERA POW
Workload
Goals
Portfolios
Analyses
Search
LMD 4.5 for
Definitions | Contact Us | Settings | Help | Back to Portal | Log Out
Saved Results at 27 Dec 16:52 from Saved Search "biofuels TX 2013".
Close
Save New
Export
Saved Searches
5 Results Found
Toggle All Fields
Genetic Mapping Of Drought Response, Nutrient Metabolism, And Other Potentially Adaptive Traits In The Bioenergy Crop Switchgrass
more like this objective...
Text Type Objective
Accession Number 0226120
Score
Genetic Mapping Of Drought Response, Nutrient Metabolism, And Other Potentially Adaptive Traits In The Bioenergy Crop Switchgrass
more like this progress...
Text Type Progress
Accession Number 0226120
Score
Genetic Mapping Of Drought Response, Nutrient Metabolism, And Other Potentially Adaptive Traits In The Bioenergy Crop Switchgrass
more like this progress...
Text Type Progress
Accession Number 0226120
Score
Building Expertise On Energy Sustainability (Bees) : An Integrative Model To Increase Research And Education In Renewable Energy Systems
more like this progress...
Text Type Progress
Accession Number 0222554
Score
A Water And Risk Management Tool For Sustainable Production Of Bioenergy Feedstocks
more like this objective...
Text Type Objective
Accession Number 1001289
Score

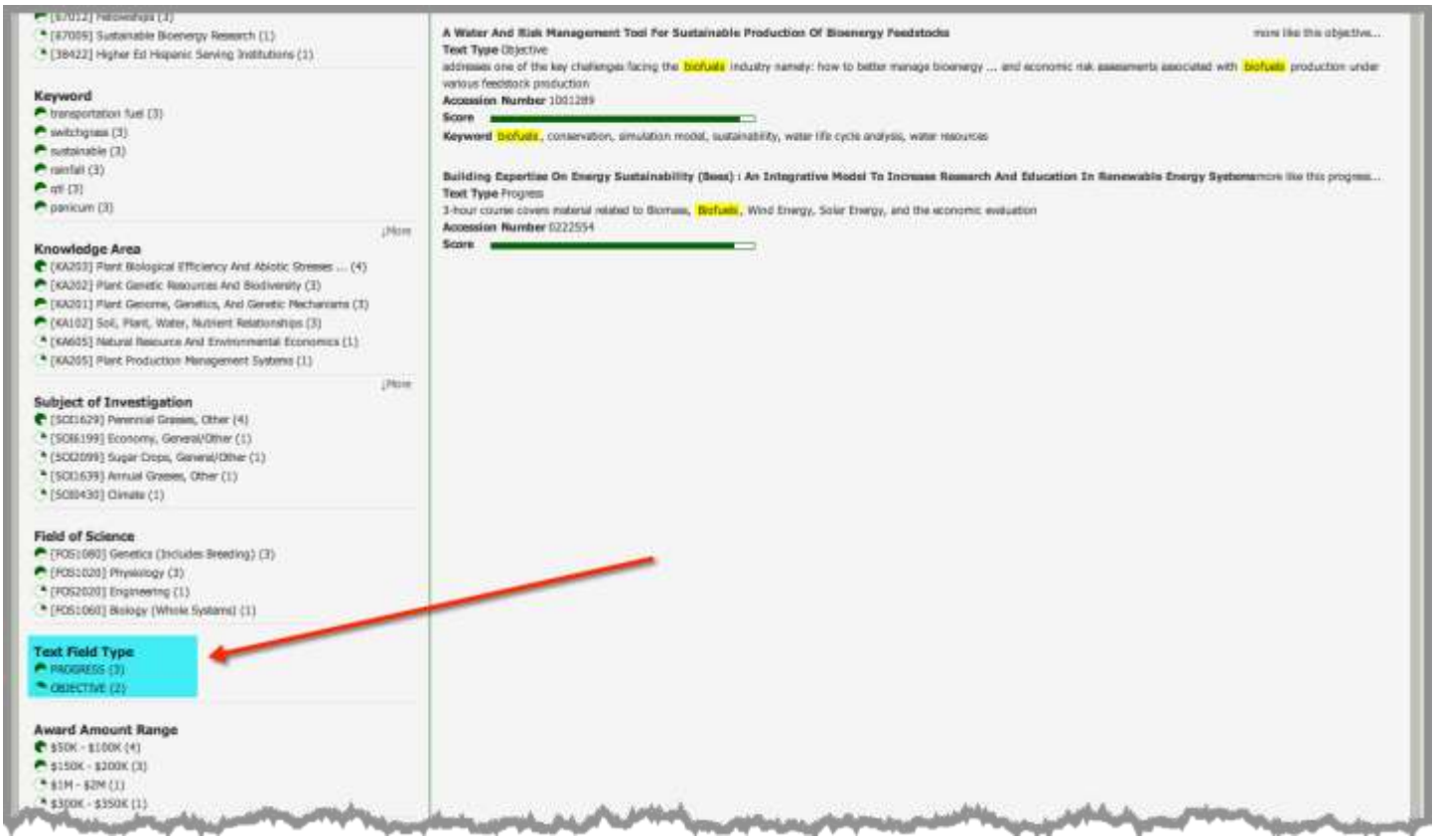
2.2.3 Exporting Results

Clicking the "Export" button will initiate and export of all of the results currently active:

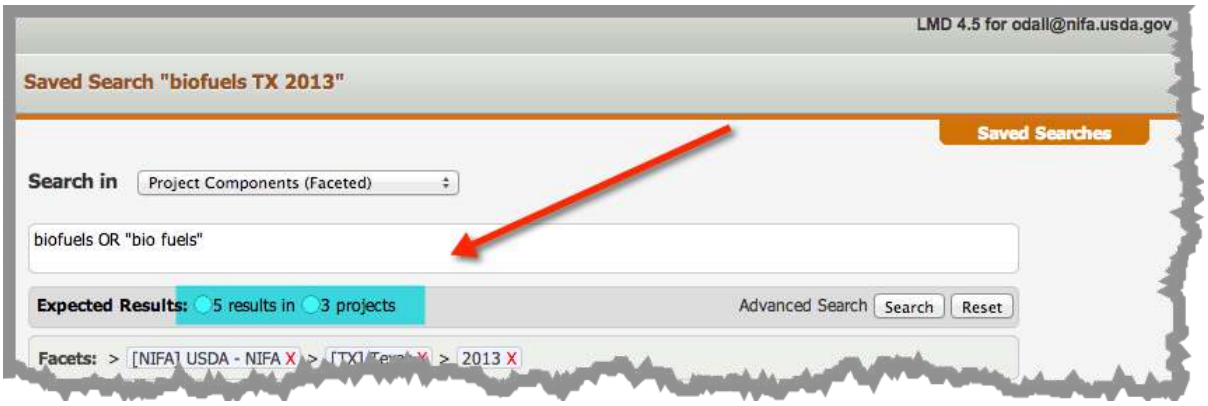
After clicking, you will be prompted to provide a name and location for the text file (in comma separated values) format:



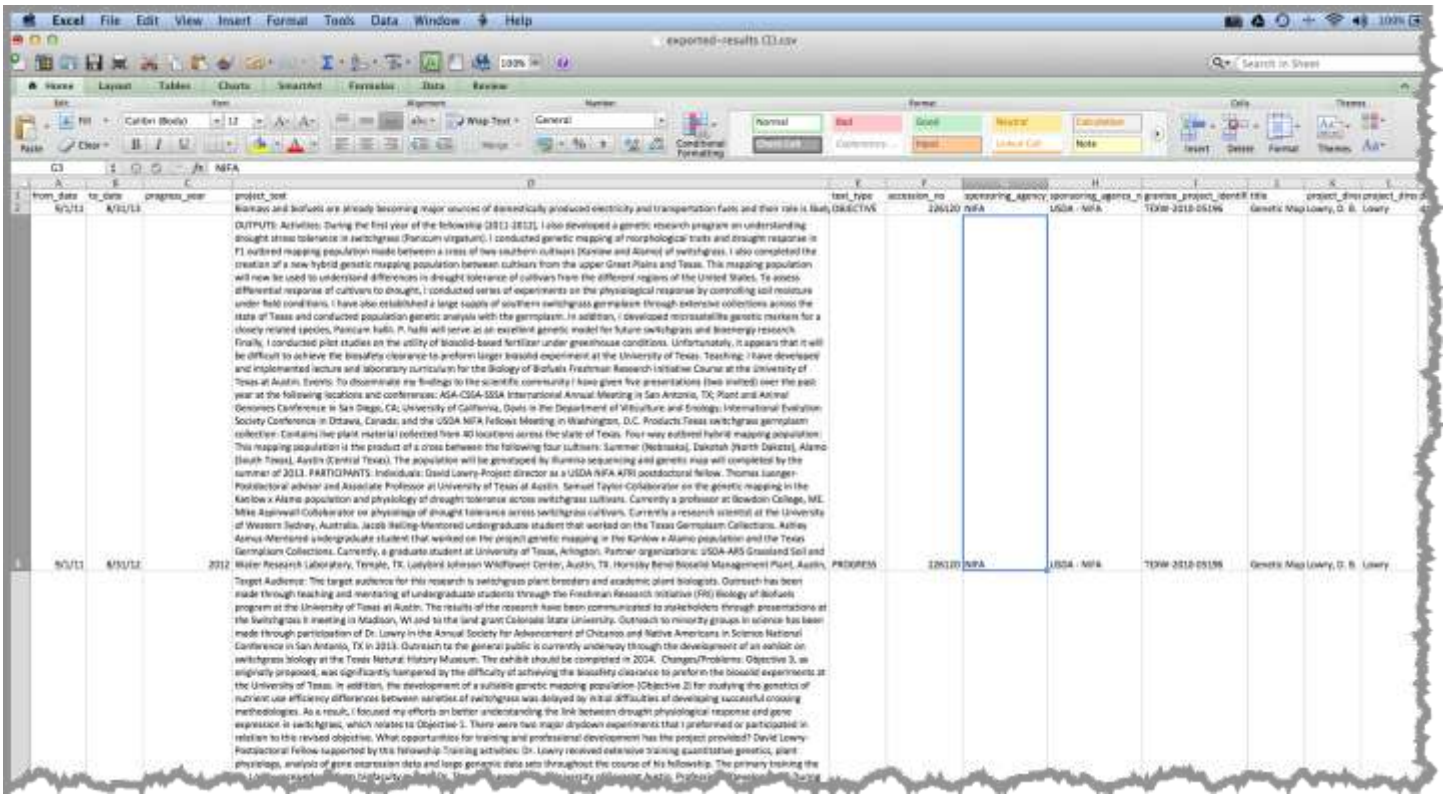
Note that there will be one row for each of the "Text Field Type" (project components) that met the criteria for the search:



In this case there are three rows of PROGRESS and two rows of OBJECTIVE among three projects that met the search criterion:



The following is a sample screen shot of the exported results in Excel:



2.3 Using Query Fields

This functionality is only available within the "Project Components (Faceted)" search domain.

2.3.1 Available Query Fields

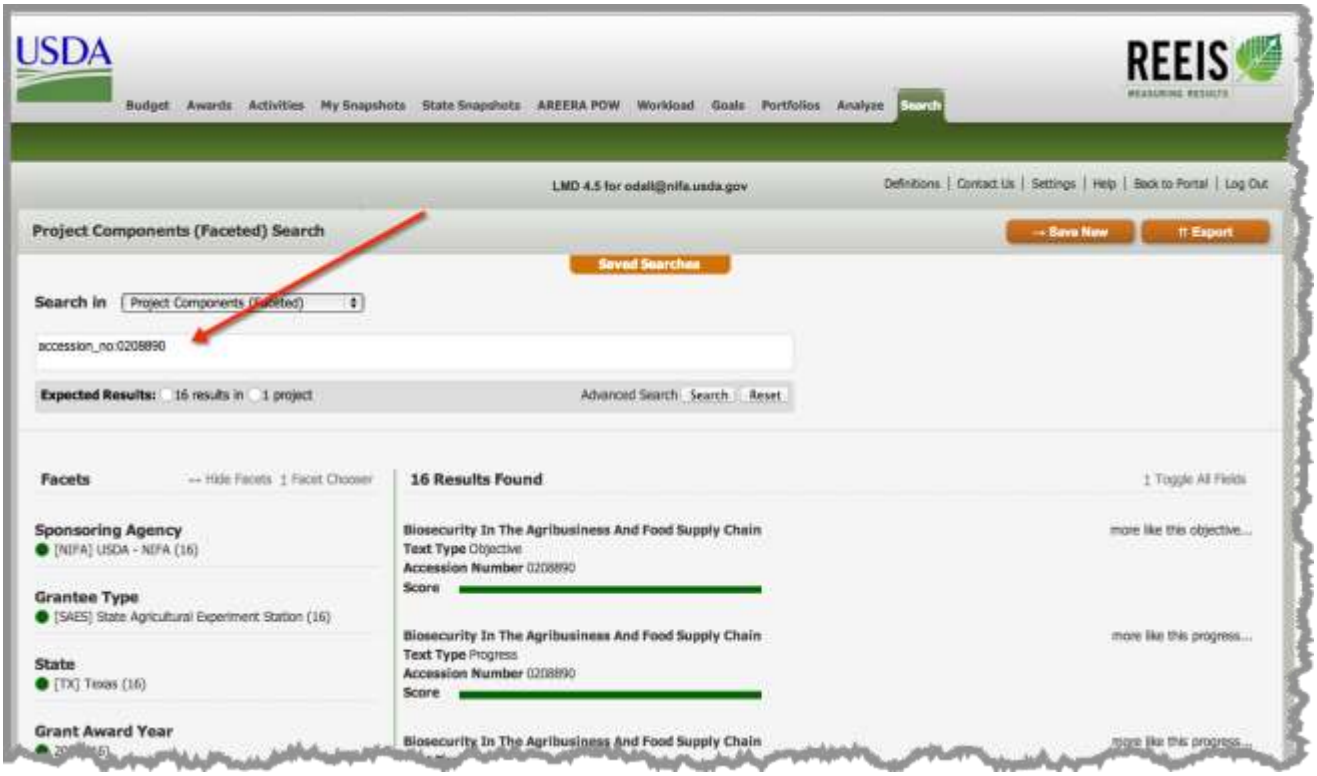
The following fields are searched by default when you enter a text search term or phrase such as "biofuels OR biomass":

FIELD NAME	DESCRIPTION	COMMENTS
from_date:	Progress Date From	
to_date:	Progress Date To	
progress_year:	Progress Year	
project_text:	Project Text	*Search only within the Objective, Progress, Impact or Publication text*
text_type:	Text Type	*"OBJECTIVE", "PROGRESS", "IMPACT","PUBLICATION" (in caps)*
accession_no:	Accession Number	**** Include the leading zero, if applicable ****
sponsoring_agency_code:	Sponsoring Agency Code	*"NIFA","ARS","ERS", "FS"*
sponsoring_agency_name:	Sponsoring Agency Name	
grantee_project_identifier:	Grantee Project Identifier	
title:	Project Title	
project_director_full_name:	Project Director (Full Name)	
project_director_last_name:	Project Director (Last Name)	
project_director_email:	Project Director Email	
project_type_code:	Project Type Code	*Project type code as per CRIS ("G", "F", etc.)*
project_type_name:	Project Type Name	*Project type name as per CRIS ("SPECIAL GRANT", "AFRI COMPETITIVE GRANT", etc.)*
status_code:	Status Code	***** Not yet implemented *****
status_name:	Status Name	***** Not yet implemented *****
research_percent:	Research Percent	***** Not yet implemented *****
education_percent:	Education Percent	***** Not yet implemented *****
extension_percent:	Extension Percent	***** Not yet implemented *****
forestry_percent:	Forestry Percent	***** Not yet implemented *****
research_basic_percent:	Research Basic Percent	
research_applied_percent:	Research Applied Percent	
research_developmental_percent:	Research Developmental Percent	
fdc:	FDC	
project_type:	Project Type	*"FORMULA" or "NON-FORMULA"*
created_at:	Text Created At	*Date Field*
updated_at:	Text Updated At	*Date Field*
grantee.name:	Grantee Name	

FIELD NAME	DESCRIPTION	COMMENTS
grantee.nifa_code:	Grantee NIFA Code	*** Not yet implemented ***
grantee.nifa_legacy_station_code:	Grantee NIFA legacy Station Code	*** "Division/Station" (DS) Code from CRIS **
grantee.duns:	Grantee Duns	
grantee.type_code:	Grantee Type Code	For example, "SAES"
grantee.type_name:	Grantee Type Name	
grantee.institution_rollup_name:	Grantee Institution Rollup Name	For example, "University of Maryland"
grantee.postal_code:	Grantee Postal Code	ZIP Code
grantee.state_code:	Grantee State Code	
grantee.state_name:	Grantee State Name	
grantee.region_code:	Grantee Region Code	
grantee.region_name:	Grantee Region Name	
grantee.country_code:	Grantee Country Code	*** Not yet implemented ***
grantee.country_name:	Grantee Country Name	*** Not yet implemented ***
grantee.address_line_1:	Grantee Addreess Line 1	
grantee.address_line_2:	Grantee Address Line 2	
grantee.address_line_3:	Grantee Address Line 3	
grantee.city_name:	Grantee City Name	In all CAPS
grantee.county_name:	Grantee County Name	In all CAPS
grantee.nifa_partner_flag:	Grantee NIFA Partner Flag	"Y" means the grantee is a NIFA Partner
grantee.type_code_with_type_name:	Grantee Type Code With Type Name	
awards.grant_year_id:	Awards Grant Year ID	This is equivalent to the numeric Fiscal Year, e.g., 2002, 2003, etc.
awards.award_amount:	Awards Award Amount	
awards.created_at:	Awards Created At	Date Field
awards.updated_at:	Awards Updated At	Date Field
kas:	Knowledge Area	List of Knowledge Areas associated with the project
sois:	Subject Of Investigation	List of Subjects of Investigation associated with the project
foss:	Field Of Science	List of Fields of Science associated with the project
keywords:	Keywords	

2.3.2 Limiting a Search By Using A Specific Query Field in Project Components (Faceted) Search

You may wish to limit the fields that are searched by the text you enter. For example, you may wish to search only the "accession_no" field when you enter text for a project accession number. Enter the field name, followed by a colon, followed by the value to be searched:



You can include any number of fields using the Boolean "AND" and "OR" operators. For example, what if you wanted to limit your search the text "biofuels" to only publications?

2.3.3 Advanced Search Capabilities

(The following text was abstracted and modified from the official Apache Lucene documentation. Apache Lucene is the underlying indexing engine used by NIFA Enterprise Search.)

Terms

A query is broken up into terms and operators. There are two types of terms: **Single Terms** and **Phrases**. A Single Term is a single word such as "test" or "hello". A Phrase is a group of words surrounded by double quotes such as "hello dolly".

Multiple terms can be combined together with Boolean operators to form a more complex query (see below).

Boolean Operators

Boolean operators allow terms to be combined through logic operators. Lucene supports AND, "+", OR, NOT and "-" as Boolean operators. (Note: Boolean operators must be ALL CAPS).

OR

NIFA Enterprise Search uses the AND operator as the default., which means entering two terms next to each other assumes the AND operator.

To search for documents that contain either "jakarta apache" or just "jakarta" use the query:

```
"jakarta apache" OR jakarta
```

AND

The AND operator matches documents where both terms exist anywhere in the text of a single document. This is equivalent to an intersection using sets. To search for documents that contain "jakarta apache" and "Apache Lucene" use the query:

```
"jakarta apache" AND "Apache Lucene"
```

NOT

The NOT operator excludes documents that contain the term after NOT. This is equivalent to a difference using sets. The symbol ! can be used in place of the word NOT. To search for documents that contain "jakarta apache" but not "Apache Lucene" use the query:

```
"jakarta apache" NOT "Apache Lucene"
```

Note: The NOT operator cannot be used with just one term. For example, the following search will return no results:

```
NOT "jakarta apache"
```

Grouping

Lucene supports using parentheses to group clauses to form sub queries. This can be very useful if you want to control the Boolean logic for a query. To search for either "jakarta" or "apache" and "website" use the query:

```
(jakarta OR apache) AND website
```

This eliminates any confusion to makes sure that the term website must exist in the results, and either term, jakarta or apache, may exist.

Wildcard Searches

Enterprise Search supports single and multiple character wildcard searches within single terms (not within phrase queries). To perform a single character wildcard search use the "?" symbol. To perform a multiple character wildcard search use the "*" symbol.

The single character wildcard search looks for terms that match the single character replaced. For example, to search for "text" or "test" you can use the search:

```
te?t
```

Multiple character wildcard searches looks for 0 or more characters. For example, to search for test, tests or tester, you can use the search:

```
test*
```

You can also use the wildcard searches in the middle of a term.

```
te*t
```

Note: You cannot use a * or ? symbol as the first character of a search.

Proximity

ES supports finding words are a within a specific distance away. To do a proximity search use the tilde, "~", symbol at the end of a Phrase. For example to search for "apache" and "jakarta" within 10 words of each other in a document use the search:

```
"jakarta apache"~10
```

Fields

When performing a search you can either specify a field, or use the default fields as outlined above. You can search any field by typing the field name followed by a colon ":" and then the term you are looking for. If you want to find the project entitled "Do it right" which contains the text "don't go this way", you can enter:

```
title:"Do it right" AND way
```

Since text is the default field, the field indicator is not required.

Note: The field is only valid for the term that it directly precedes, so the query

```
title:Do it right
```

will only find "Do" in the title field. It will find "it" and "right" in the default fields. Note: To search a complete title phrase, see the **Field Grouping** instructions below to make the phrase into a single term.

Range Searches

Range Queries allow one to match documents whose field(s) values are between the lower and upper bound specified by the Range Query. Range Queries can be inclusive or exclusive of the upper and lower bounds. Sorting is done lexicographically.

```
to_date:[20020101 TO 20030101]
```

This will find documents whose to_date field has values between 20020101 and 20030101, inclusive. To exclude the two defined dates from the list, use curly brackets (braces) instead:

```
to_date:{20020101 TO 20030101}
```

Note that Range Queries are not reserved for date fields. You could also use range queries with non-date fields:

```
title:{Aida TO Carmen}
```

This will find all documents whose titles are between Aida and Carmen, but not including Aida and Carmen.

Remember, inclusive range queries are denoted by square brackets. Exclusive range queries are denoted by curly brackets, also known as “braces.”

Boosting a Term

Lucene provides the relevance level of matching documents based on the terms found. To boost a term use the caret, "^", symbol with a boost factor (a number) at the end of the term you are searching. The higher the boost factor, the more relevant the term will be. Boosting allows you to control the relevance of a document by boosting its term. For example, if you are searching for jakarta apache and you want the term "jakarta" to be more relevant boost it using the ^ symbol along with the boost factor next to the term. You would type:

```
jakarta^4 apache
```

This will make documents with the term jakarta appear more relevant. You can also boost Phrase Terms as in the example:

```
"jakarta apache"^4 "Apache Lucene"
```

By default, the boost factor is 1. Although the boost factor must be positive, it can be less than 1 (e.g. 0.2).

Field Grouping

Lucene supports using parentheses to group multiple clauses to a single field. To search for a title that contains both the word "return" and the phrase "pink panther" use the query:

```
title:(+return +"pink panther")
```